Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 9228–9232, September 1994
Biochemistry

Construction and characterization of a normalized cDNA library

(brain mRNA/DNA circles/reassociation kinetics)

Marcelo Bento Soares*†, Maria de Fatima Bonaldo*†, Pierre Jelene*†, Long Su*†, Lee Lawton*†, and Argiris Efstratiadis‡

Departments of *Psychiatry and ‡Genetics and Development, Columbia University, and †The New York State Psychiatric Institute, 722 West 168th Street, New York, NY 10032

Communicated by Isidore S. Edelman, June 8, 1994

ABSTRACT **We have developed a simple procedure based on reassociation kinetics that can reduce effectively the high variation in abundance among the clones of a cDNA library that represent individual mRNA species. For this normalization, we used as a model system a library of human infant brain cDNAs that were cloned directionally into a phagemid vector and, thus, could be easily converted into single-stranded circles. After controlled primer extension to synthesize a short complementary strand on each circular template, melting and reannealing of the partial duplexes at relatively low C_0t, and hydroxyapatite column chromatography, unreassociated circles were recovered from the flow through fraction and electroporated into bacteria, to propagate a normalized library without a requirement for subcloning steps. An evaluation of the extent of normalization has indicated that, from an extreme range of abundance of 4 orders of magnitude in the original library, the frequency of occurrence of any clone examined in the normalized library was brought within the narrow range of only 1 order of magnitude.**

The mRNAs of a typical somatic cell are distributed in three frequency classes (1, 2) that are presumably maintained in representative cDNA libraries. The classes at the two extremes (*ca.* 10% and 40–45% of the total, respectively) include members occurring at vastly different relative frequencies. On average, the most prevalent class consists of about 10 mRNA species, each represented by 5000 copies per cell, whereas the class of high complexity comprises 15,000 different species each represented by 1–15 copies only. Rare mRNAs are even more under represented in the brain, a tissue exhibiting an exceptionally high sequence complexity of transcripts (3–5).

Although even the rarest mRNA sequence from any tissue is likely to be represented in a cDNA library of 10^7 recombinants, its identification is very difficult (its frequency of occurrence may be as low as 2×10^{-6} on average or even 10^{-7} for complex tissues such as the brain). Thus, for a variety of purposes, it is advantageous to apply a normalization procedure and bring the frequency of each clone in a cDNA library within a narrow range (generation of a perfectly equimolar cDNA library is practically impossible in our experience). Normalized cDNA libraries can facilitate positional cloning projects aiming at the identification of disease genes, can increase the efficiency of subtractive hybridization procedures, and can significantly facilitate genomic research pursuing chromosomal assignment of expressed sequences and their localization in large fragments of cloned genomic DNA (exon mapping). Normalization makes feasible the gridding of cDNA libraries on filters at high density by reducing the number of clones to be arrayed (gridding 10^7 clones for 1× coverage of a non-normalized library is not a feasible task). Finally, by increasing the frequency of occurrence of rare cDNA clones while decreasing simultaneously the percentage of abundant cDNAs, normalization can expedite significantly the development of expressed sequence databases by random sequencing of cDNAs.

Although cDNA library normalization could be achieved by saturation hybridization to genomic DNA (6), this approach is impractical, since it would be extremely difficult to provide saturating amounts of the rarer cDNA species to the hybridization reaction. The alternative is the use of reassociation kinetics: assuming that cDNA reannealing follows second-order kinetics, rarer species will anneal less rapidly and the remaining single-stranded fraction of cDNA will become progressively normalized during the course of the reaction (6–8). As we report here, we have used this kinetic principle to develop a method for normalization of a directionally cloned cDNA library that has significant advantages over two previously reported similar procedures (refs. 7 and 8; see *Results and Discussion*).

MATERIALS AND METHODS

cDNA Library Construction. Poly(A)$^+$ RNA isolated from the entire brain of a female infant (72 days old), who died in consequence of spinal muscular atrophy, was used for construction of a cDNA library (IB) as described (9, 10). As a primer for first-strand cDNA synthesis, we used the oligonucleotide 5′-AACTGGAAGAATTCGCGGCCGCAGGAAT$_{18}$-3′, which contains a *Not* I site (underlined). After ligation to *Hin*dIII adaptors, the cDNAs were digested with *Not* I and cloned directionally into the *Hin*dIII and *Not* I sites of a phagemid vector (L-BA) constructed by modifying pEMBL-9(+) (11). L-BA carries an ampicillin-resistance gene, plasmid and filamentous phage (f1) origins of replication, and cloning sites (5′ *Hin*dIII–*Bam*HI–*Not* I–*Eco*RI 3′). Superinfection of bacteria with the helper phage M13K07 (12) converts duplex plasmids into single-stranded DNA circles containing message-like strands of the cDNA inserts.

Preparation of Single-Stranded Library DNA. Plasmid DNA from the IB library was electroporated into *Escherichia coli* DH5α F′ bacteria, and the culture was grown under ampicillin selection at 37°C to an OD_{600} of 0.2, superinfected with a 20-fold excess of the helper phage M13K07, and harvested after 4 hr for preparation of single-stranded plasmids, as described (12). To eliminate contaminating double-stranded, replicative form (RF) DNA, 20 μg of the preparation was digested with *Pvu*II (which cleaves only duplex DNA molecules), extracted with phenol/chloroform, diluted by addition of 2 ml of loading buffer (0.12 M sodium phosphate buffer, pH 6.8/10 mM EDTA/1% SDS), and purified by hydroxyapatite (HAP) chromatography at 60°C, using a column preequilibrated with the same buffer (1-ml bed volume; 0.4 g of HAP). After a 6-ml wash with loading buffer,

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: HAP, hydroxyapatite.

this volume was combined with the flow through fraction, and the sample was extracted twice with water-saturated 2-butanol, once with dry 2-butanol, and once with water-saturated ether (3 volumes per extraction). The sample was desalted by passage through a Nensorb column (DuPont/NEN) according to the manufacturer's specifications, concentrated by ethanol precipitation, and electrophoresed in a low-melting agarose gel to remove helper phage DNA and any residual tRNA contaminant or oligoribonucleotides (breakdown products from the RNase A digestion used during purification). The region of the gel containing the single-stranded library was excised and, after β-agarase (New England Biolabs) digestion, the DNA was purified and ethanol-precipitated.

cDNA Library Normalization. The IB cDNA library was normalized (see Fig. 1) in two consecutive rounds to derive the normalized libraries ^{1}NIB and ^{2}NIB, by using the following procedure. To synthesize a partial second strand of about 200 nt by limited extension, 9 pmol of the oligonucleotide primer 5'-GGCCGCAGGAAT$_{15}$-3' was added to 4.5 pmol of single-stranded IB library DNA in a 150-μl reaction mixture containing 30 mM Tris·HCl (pH 7.5); 50 mM NaCl; 15 mM $MgCl_2$; 1 mM dithiothreitol; 0.1 mM dNTPs; 2.5 mM ddATP, ddCTP, and ddGTP; and a trace of [α-^{32}P]dCTP. The mixture was incubated for 5 min at 60°C and for 15 min at 50°C, the temperature was lowered to 37°C, 75 units of Klenow DNA polymerase (United States Biochemical) was added, and the incubation was continued for 30 min. The reaction was terminated by addition of EDTA (20 mM), extracted with phenol/chloroform, diluted with 2 ml of HAP loading buffer containing 50 μg of sonicated and denatured salmon sperm DNA carrier, and chromatographed on HAP, as described above. After washing, the partial duplex circles bound to HAP were eluted from the column with 6 ml of 0.4 M phosphate buffer, pH 6.8/10 mM EDTA/1% SDS. The concentration of phosphate in the eluate was lowered to 0.12 M by addition of 14 ml of water containing 50 μg of DNA carrier, and the chromatographic step was repeated. The final eluate was extracted and desalted as described above and the DNA was ethanol-precipitated. The pellet (112 ng) was dissolved in 2.5 μl of formamide and the sample was heated for 3 min at 80°C under a drop of mineral oil to dissociate the DNA strands. For an annealing reaction, the volume was brought to 5 μl by adding 0.5 μl of 0.1 M Tris·HCl, (pH 7.5) containing 0.1 M EDTA, 0.5 μl of 5 M NaCl, 1 μl (5 μg) of $(dT)_{25-30}$, and 0.5 μl (0.5 μg) of the extension primer. The last two ingredients were added to block stretches of adenine residues [representing the initial poly(A) tails] and regions complementary to the oligonucleotide on the single-stranded DNA circles. The annealing mixture was incubated at 42°C, and a 0.5-μl aliquot was withdrawn at 13 hr (calculated C_0t, 5.5). The unhybridized single-stranded circles (normalized library) were separated from the reassociated partial duplexes by HAP chromatography and then recovered from the flow through fraction as described above. Since we, and others (13), have observed that the electroporation efficiency of partially repaired circular molecules is increased by about 100-fold in comparison with single-stranded circles, the normalized cDNA circles were converted to partial duplexes by primer extension using random hexamers and T7 DNA polymerase (Sequenase version II; United States Biochemical), in a 10–20 μl reaction mixture containing 1 mM dNTPs. After addition of EDTA to 20 mM, phenol extraction, and ethanol precipitation, the cDNAs were dissolved in 10 mM Tris·HCl, pH 7.5/1 mM EDTA, and electroporated into competent bacteria (DH10B; GIBCO/BRL). To determine the number of transformants, 1 hr after the electroporation a 10-μl aliquot of the culture was plated on an LB agar plate containing 75 μg/ml ampicillin (extrapolation from these data indicated that a normalized library of 2.5 × 10^6 colonies was obtained). Supercoiled plasmid DNA was then prepared (^{1}NIB library) with a Qiagen plasmid kit (Qiagen, Chatsworth, CA). The same protocol was used for a second round of normalization (calculated C_0t, 2.5) to derive the ^{2}NIB library (1.3 × 10^7 transformants) from a preparation of ^{1}NIB single-stranded circles, except that the HAP purification step after primer extension to synthesize short complementary strands was omitted.

Colony Hybridization. For screening, colonies were grown on duplicate nylon filters (GeneScreen*Plus*; DuPont/NEN) that were processed as described (14) and hybridized at 42°C in 50% formamide/5× Denhardt's solution/0.75 M NaCl/0.15 M Tris·HCl, pH 7.5/0.1 M sodium phosphate/0.1% sodium pyrophosphate/2% SDS containing sheared and denatured salmon sperm DNA at 100 μg/ml. Radioactive probes were prepared by random primed synthesis (15, 16) using the Prime-it II kit (Stratagene).

DNA Sequencing. Double-stranded plasmid DNA templates were prepared by using the Wizard Minipreps DNA purification system (Promega) and sequenced from both ends by using the universal forward and reverse M13 fluorescent primers. Reactions were assembled on a Biomek 1000 workstation (Beckman) and then transferred to a thermocycler (Perkin–Elmer/Cetus) for cycle sequencing. Reaction products were analyzed on an automated 370A DNA sequencer (Applied Biosystems). Nucleic acid and protein database searches were performed at the National Center for Biotechnology Information server using the BLAST algorithm (17).

RESULTS AND DISCUSSION

Experimental Strategy. To develop a normalization procedure, shown schematically in Fig. 1, and at the same time

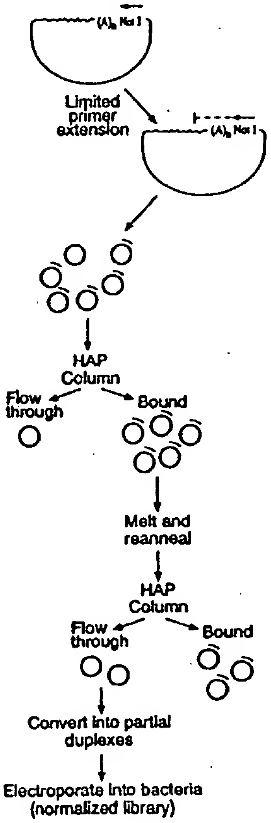


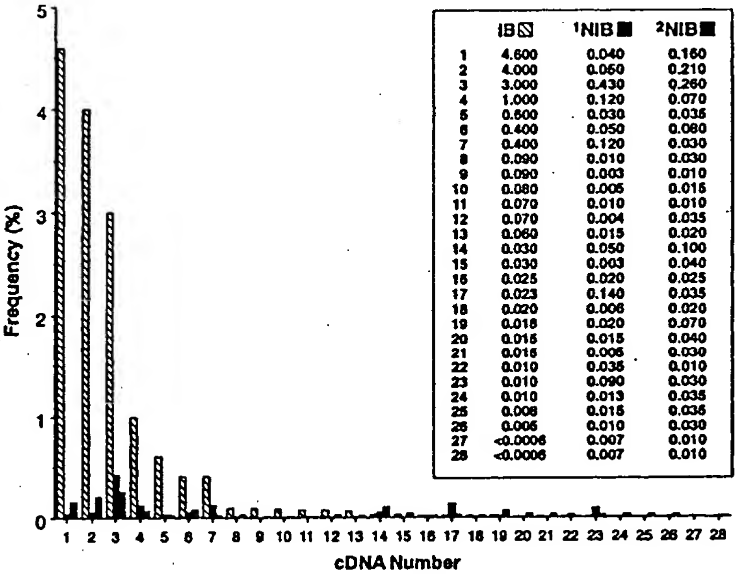
FIG. 1. Diagram of the normalization procedure. Single-stranded circles of a library of directionally cloned cDNAs are primer extended under controlled conditions to generate complementary strands of about 200 ± 20 nt, and the resulting partial duplexes are purified from unprimed circles by HAP chromatography. Bound DNA is melted and reannealed to a relatively low C_0t (see text). The remaining single-stranded circles (normalized library) are isolated by HAP chromatography, converted into partial duplexes by random priming, and electroporated into bacteria for amplification.

increase the utility of the normalized model cDNA library, we first constructed a high-quality brain cDNA library (IB) that has the following features (10): the average size of a cDNA insert is 1.7 kb, often providing coding-region information by sequencing from the 5′ end; the length of the segment representing the mRNA poly(A) tail is short, allowing an increase in the output of useful sequencing information from the 3′ end; the frequency of nonrecombinant clones is extremely low (0.1%); and chimeric cDNAs have not been encountered, after single-pass sequencing of >2000 clones (10, 18). However, the latter analysis also demonstrated that 13% of the clones in the IB library lacked poly(A) tails and were presumably derived from aberrant priming.

To preserve the length of the cDNAs, avoid differential loss of sequences, and alleviate a need for subcloning steps after normalization, we excluded from our protocol the use of PCR and chose directional cloning into a phagemid vector. Such vectors have been previously used advantageously for cDNA library subtractions (13), although normalization was not attempted. This cloning regime readily provides single strands that can be used both for annealing and for direct propagation in bacteria. In control experiments (data not shown), we assessed the frequency of occurrence of abundant cDNAs (encoding α- and β-tubulin, elongation factor 1α, and myelin basic protein) and demonstrated that, at least by this criterion, the representation of clones in the starting library remained unchanged after conversion into single-stranded circles. We also note that electrophoretic purification of the circles prior to use is necessary, to remove contaminating oligoribonucleotides (see *Materials and Methods*), whose presence would result in undesirable internal priming events during the first step of our protocol.

In contrast with our scheme, two other PCR-based normalization methods (7, 8) necessitate the use of subcloning steps. In one of these approaches (7), sheared cDNAs (0.2–0.4 kb) were ligated to a linker–primer, amplified by PCR, normalized kinetically, reamplified, and finally cloned directionally in such a way that only 3′-terminal sequences (almost exclusively 3′ noncoding regions) were purposely preserved. The steps of the second scheme (8) were similar, except that the process started from cloned, randomly primed, and relatively short cDNAs, initially selected to minimize length-dependent differential PCR amplification. Thus, both coding and noncoding regions were represented in the final normalized library, but in pieces.

While maintaining length and representation of mRNA regions, our protocol (Fig. 1) also addresses successfully the problem recognized in the first of the alternative approaches (7). It was considered that the 3′ noncoding region is almost always unique to the transcript that it represents and is expected, therefore, to anneal only to its complement. In contrast, cross-hybridization of coding regions belonging to unequally represented members of oligo- or multigene families could result in the elimination of rarer members from the population during the normalization process. This possibility is precluded in our method, which begins with the synthesis, from the 3′ end of the cDNA, of a short complementary strand on the circular single-stranded cDNA template under controlled conditions, calibrated to yield strands with a narrow size distribution (200 ± 20 nt). Since the average length of 3′ noncoding regions in brain mRNAs is 750 nt (19), the vast majority of synthesized complementary strands participating in the annealing reaction should be devoid of coding region sequences. However, after this partial extension step, purification of the products by HAP chromatog-



	IB ▨	¹NIB ■	²NIB ■
1	4.600	0.040	0.160
2	4.000	0.060	0.210
3	3.000	0.430	0.260
4	1.000	0.120	0.070
5	0.600	0.030	0.035
6	0.400	0.050	0.080
7	0.400	0.120	0.030
8	0.090	0.010	0.030
9	0.090	0.003	0.010
10	0.080	0.005	0.015
11	0.070	0.010	0.010
12	0.070	0.004	0.035
13	0.060	0.015	0.020
14	0.030	0.050	0.100
15	0.030	0.003	0.040
16	0.025	0.020	0.025
17	0.023	0.140	0.035
18	0.020	0.006	0.020
19	0.018	0.020	0.070
20	0.015	0.015	0.040
21	0.015	0.006	0.030
22	0.010	0.035	0.010
23	0.010	0.090	0.030
24	0.010	0.013	0.035
25	0.008	0.015	0.035
26	0.006	0.010	0.030
27	<0.0006	0.007	0.010
28	<0.0006	0.007	0.010

FIG. 2. Comparison of the frequencies of cDNA probes in the original (IB) and two normalized (¹NIB and ²NIB) libraries. The indicated percentages of 28 cDNA sequences in the three libraries, tabulated in order of decreasing frequency in the IB library, are shown in the form of a histogram to visualize normalization. Frequencies were calculated from the number of positive colonies after hybridization of duplicate filters containing 500–180,000 colonies from each of the three cDNA libraries with the following 28 probes: 1, elongation factor 1α; 2, α-tubulin; 3, β-tubulin; 4, myelin basic protein; 5, aldolase; 6, 89-kDa heat shock protein; 7, γ-actin; 8, secretogranin; 9, microtubule-associated protein; 11, vimentin; 13, a cDNA randomly picked from the ¹NIB library similar to a mouse cysteine-rich intestinal protein (¹NIB-2, GenBank accession nos. T09996 and T09997); 19, a cDNA isolated from the ¹NIB library homologous to the human endogenous retrovirus RTVLH2 (cDNA-20, accession nos. L13822 and L13823); 20, histone H2b.1; 23, a cDNA randomly picked from the ¹NIB library encoding the human polyposis (*DP1* gene) mRNA (¹NIB-227, accession nos. T10266 and T10267); 27, a cDNA randomly picked from the ¹NIB library related to the human endogenous retrovirus ERV9 gene (¹NIB-114, accession nos. T10086 and T10087); the remaining brain cDNAs are novel, and except for nos. 10, 18, 21, and 25, they were randomly picked from the ¹NIB library.

raphy is necessary to eliminate single strands of the IB library lacking poly(A) tails that cannot participate in primed synthesis. We repeat the chromatographic step to reduce the background to negligible levels, since after the first passage through the HAP column about 0.1% of pure single strands bind nonspecifically. However, during the second round of normalization to derive the ^{2}NIB library, we omitted this step since we showed that 187 clones, which were picked randomly and sequenced from the ^{1}NIB library (see below), all contained 3′ poly(A) stretches. The remaining steps of our procedure entail melting and reannealing of the partial duplexes, followed by purification of unreassociated circles (normalized library) by HAP chromatography and electroporation into bacteria (Fig. 1).

Characterization of Normalized cDNA Libraries. To evaluate the extent of normalization achieved with our method, we compared the IB, ^{1}NIB, and ^{2}NIB libraries by colony hybridization. For this analysis, we used 28 cDNA probes chosen to represent various frequencies of occurrence within a wide range (at least 4 orders of magnitude: 4.6% to <0.0006%) in the IB library (Fig. 2). However, an additional comparison of these results with independent theoretical estimates was necessary, to provide a further assessment of the degree of normalization, especially because the ^{1}NIB library was derived after incubation to a relatively low C_0t (5.5) during the reannealing step of our procedure. When relatively high C_0t values were used in our initial attempts to normalize the IB library, we obtained unsatisfactory results (high background) that we attribute to technical problems inherent to the procedure. Nevertheless, a reevaluation of brain cDNA hybridization data (ref. 20; see Table 1) suggests that a relatively low C_0t would suffice for our purpose, to bring the frequency of each library clone within a narrow range.

For our calculations (Table 1), which should be regarded as rough but indicative estimates, we used a set of reliable hybridization data that are available only for mouse brain mRNAs (20), assuming that these measurements should not differ significantly among mammals (in all cases examined, including humans, the average amount of RNA per brain cell and the number of cells per gram of tissue are practically the same; see, e.g., refs. 29 and 30). These calculations show that at C_0t 5.5, of the three kinetic classes of mRNAs, the most abundant species are drastically diminished, while all frequencies are brought within the range of 1 order of magnitude (Table 1, compare columns b and h and columns f and i). Our experimental results (Fig. 2) show that the same range was achieved after a single round of normalization at this C_0t (5.5). Thus, for all practical purposes, a single cycle is probably sufficient. Secondary normalization (calculated C_0t = 2.5) to derive the ^{2}NIB library, although it did not result in a dramatic improvement, preserved the range of frequencies, while making the differences among individual sequences narrower overall (Fig. 2). Eleven of the 28 probes used in this analysis were derived from clones that were randomly picked from the ^{1}NIB library. The overall frequency fold variation was reduced from >7667 (4.6/<0.0006) in the IB library to 133 (0.4/0.003) and 26 (0.1/0.01) in the ^{1}NIB and ^{2}NIB libraries, respectively. However, some unexplained anomalies were also observed for a small minority of clones, whose already reduced frequencies in the ^{1}NIB library were somewhat increased in the ^{2}NIB library (Fig. 2).

To provide a further indication that normalization was successful, we sequenced from both ends 187 cDNA clones that were randomly picked from the ^{1}NIB library (GenBank accession numbers T09994–T10011 and T10014–T10369). With the exception of 4 clones, which carried sequences corresponding to human mitochondrial 16S rRNA, all other cDNAs of this pool were unique, in agreement with the expectation for a normalized library. To further investigate the effect of the normalization procedure on the subset of mitochondrial 16S rRNA clones (1.4%, 1%, and 0.4% in the IB, ^{1}NIB and ^{2}NIB libraries, respectively), we compared the sequences of a number of 16S rRNA clones isolated from both the IB and ^{1}NIB libraries (kindly provided by M. Adams, Institute for Genomic Research and J. Sikela, University of Colorado). This analysis (data not shown) revealed that the 16S rRNA clones isolated from ^{1}NIB did not correspond to

Table 1. Estimates of frequencies of brain mRNAs

Component[a]	%[b]	k_{pfo} (pure)[c]	Complexity,[d] kb	No. of RNA species[e]	Frequency per species,[f] %	k_{so}[g]	Component at C_0t 5.5,[h] %	Final frequency per species,[i] %
I	16	10	96	36	0.44	6.15	0.7	0.02
II	46	0.165	5,800	2,150	0.02	0.10	44.2	0.02
III	38	0.0079	122,000	45,000	0.0008	0.0048	55.1	0.0012

[a] The experimental data of pseudo-first-order hybridization kinetics of cDNA tracer, which was synthesized from mouse brain poly(A)$^{+}$ polysomal mRNA and driven by its template (20), were solved by computer (unconstrained fit) into three kinetic components, using the EXCESS function of a least-squares curve-fitting program (21).

[b] The fraction of total occupied by each of the components is shown, after a minor correction (at completion, practically all of the tracer had reacted). These numbers (and all other numbers) in the table have been rounded.

[c] The computer-calculated pseudo-first-order hybridization rate constant (k_{pfo}; $M^{-1}\cdot sec^{-1}$) for each component was divided by each of the values in column b, to derive k_{pfo} (pure).

[d] The complexity (i.e., length of unique sequence) was calculated by considering the data from a calibration kinetic standard: cDNA synthesized from encephalomyocarditis virus RNA (complexity, 9.7 kb) that was driven by its template [k_{pfo} (pure), 99]. Thus, each of the values in column d is the ratio $(99 \times 9.7)/k_{pfo}$ (pure). The complexity calculated for the rarest component (III) matches closely the values obtained from additional kinetic experiments using cDNA enriched for infrequent sequences (22, 23) and also the data of saturation experiments with single-copy genomic DNA tracer (24, 25).

[e] The number of different RNA species in each component was estimated from their complexities by assuming that the average size of brain mRNA is 2.7 kb (26). A conjecture (26) that rare brain mRNAs are longer than this value (hypothetically 5 kb on average) has not been supported by hard evidence.

[f] The initial average frequency of an individual mRNA species of each component in the entire population of mRNA molecules is the ratio of values in column b to those in column e.

[g] To assess the behavior of these kinetic components under the annealing conditions that we used for normalization (C_0t, 5.5; length of complementary sequence in annealing strands, 0.2 kb), we first calculated the second-order reassociation rate constant (k_{so}; $M^{-1}\cdot sec^{-1}$) for each component. For this calculation, we considered that the k_{so} of a single and pure kinetic component with a complexity of 1 kb reacting at a fragment length of 0.2 kb is 590 (27, 28). Thus each k_{so} value is 590 divided by the complexity in column d.

[h] To determine the percentage of the leftover of each component in the population at C_0t 5.5, we first used the k_{so} values in column g to calculate the fraction remaining single-stranded, according to the equation $C/C_0 = 1/(1 + kC_0t)$ and then normalized the derived values to a total of 100%.

[i] The final average frequency of an individual mRNA species of each component is the ratio of values in column h to those in column e.

the predominant 16S rRNA species present in the IB library. Interestingly, in 17 of 19 16S rRNA clones sequenced from the IB library, the position of the A tract was the same as that present in the mature 16S rRNA. In contrast, all 8 clones sequenced from the ^{1}NIB library represented truncated versions of the 16S rRNA, in which different lengths of the 3′ terminal sequence were absent. Such truncated clones are under represented in the IB library (2 of 19). Therefore, their frequency was increased by normalization, as expected, while the 16S rRNA clones of the most prevalent form were reduced. It is likely that the shorter clones represent bona fide copies of naturally occurring truncated 16S rRNA molecules (ref. 31–33; to be discussed elsewhere).

Database searches (both BLASTN and BLASTX; ref. 17) revealed that of the 183 cDNAs examined, 152 (83%) were unknown (no hits), 15 (8.2%) corresponded to known human sequences, 5 (2.7%) were novel but related to known human sequences, 4 (2.2%) were homologous to mammalian sequences, and 7 (3.8%) were homologous to known sequences from various nonmammalian organisms.

In contrast to these results, when 1633 randomly picked clones from the non-normalized IB library were sequenced mostly (88%) from the 5′ end, the percentage of unknown sequences was significantly lower than in our case (63%), while about 30% of the clones were sequenced twice or more (up to 50) times (10). Similar results were obtained by sequencing 493 random IB clones exclusively from the 3′ end (18). Of the initially abundant cDNAs, which were sequenced multiple times in both of these studies, those encoding elongation factor 1α, α-tubulin, β-tubulin, myelin basic protein, and γ-actin (corresponding to our probes 1–4 and 7; Fig. 2) were absent from the pool of 187 clones that we examined. Moreover, only 15 of the unique 183 clones that we sequenced from the ^{1}NIB library (8%) had been previously identified in the collection of the sequenced 1633 IB clones.

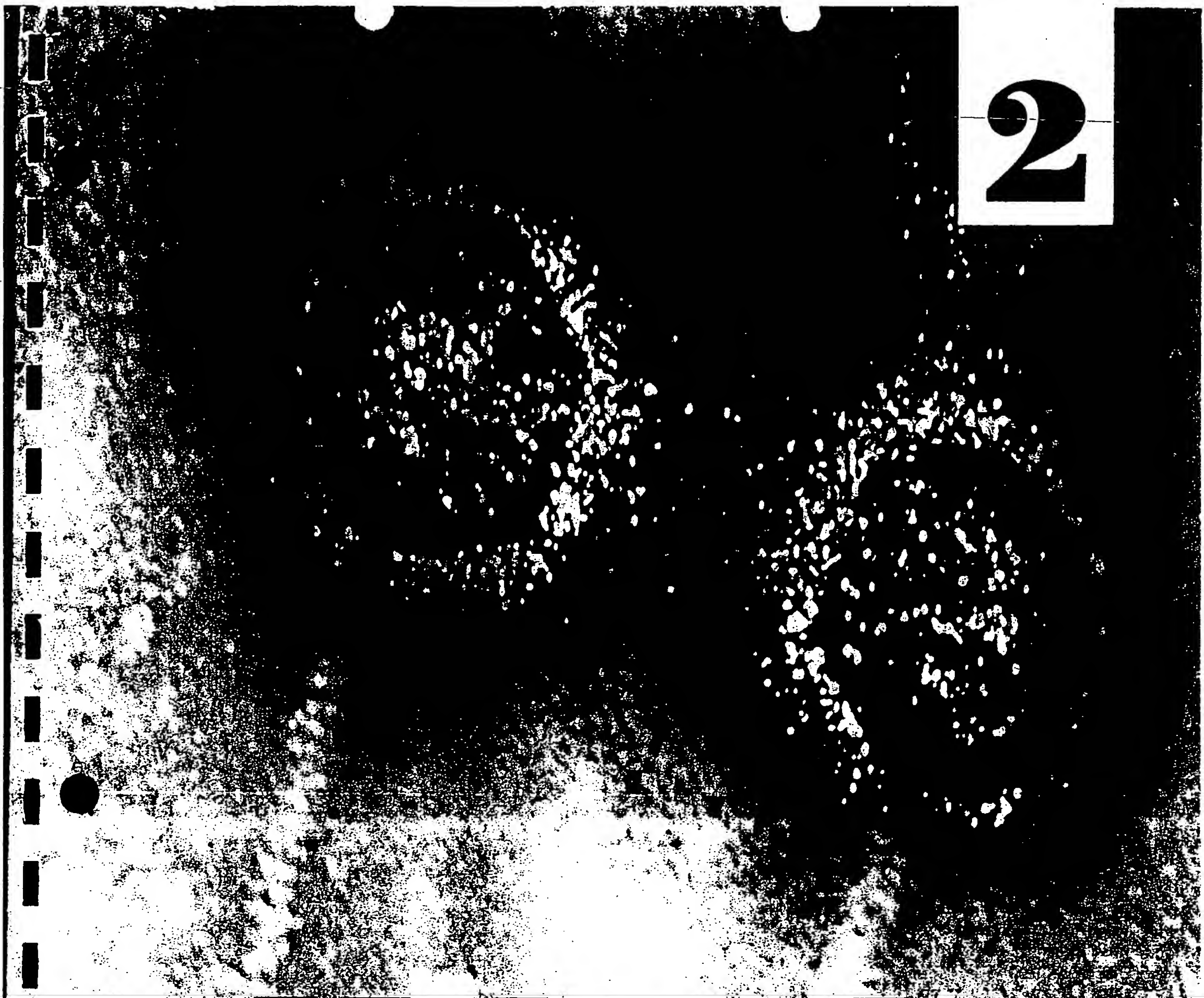
Eighteen of the unknown cDNAs that we sequenced (10% of the total clones) carried *Alu* repetitive elements (6 at the 5′ end; 11 at the 3′ end; and 1 at both ends). Thus, as previously observed (8), the frequency of cDNAs containing *Alu* repeats is not reduced by normalization. This phenomenon can be attributed to sequence heterogeneity among *Alu* family members, which are able to form imperfect hybrids that probably cannot bind to HAP. However, this is not a disadvantageous property, since it prevents elimination of rare *Alu*-carrying cDNAs from the population.

To assess whether the normalization procedure had skewed the distribution of lengths favoring shorter cDNA clones, Southern blots of released inserts from the IB, ^{1}NIB, and ^{2}NIB plasmids were hybridized with several of the cDNA probes used in Fig. 2 individually. The results (not shown) demonstrated that the intensity of hybridization signals varied as expected, but the size of each hybridizing fragment remained the same.

Note. Sasaki *et al.* (34) have described an alternative normalization procedure, in which a cDNA library was constructed following depletion of abundant mRNA species by sequential cycles of hybridization to matrix-bound cDNA. However, this procedure does not seem to be more advantageous than ours, while its actual practical potential remains to be assessed, as the putative normalized library was not adequately characterized.

We thank Greg Lennon (Lawrence Livermore National Laboratories) for arraying 40,000 clones of the ^{1}NIB library and making it available to other investigators; M. Adams and J. Sikela for kindly providing sequences of 16S rRNA clones isolated from the IB and ^{1}NIB libraries, respectively; M. Adams for making a set of 1633 sequences available to us prior to publication; and Rachel Yarmolinsky for help with the figures. This work was supported by Grant FG02-91ER61233 from the U.S. Department of Energy and by Grants 1R55HD28422 and BRSG 2 S07 RR05395-30 from the National Institutes of Health to M.B.S.; by Grants HG00362, HG00424, and HG00861 from the National Institutes of Health to A.E.; and by a generous gift from the W. M. Keck Foundation.

1. Davidson, E. H. & Britten, R. J. (1979) *Science* **204**, 1052–1059.
2. Bishop, J. O., Morton, J. G., Rosbash, M. & Richardson, M. (1974) *Nature (London)* **250**, 199–204.
3. Hahn, W. E. & Owens, G. P. (1988) in *The Molecular Biology of Neurological Disease*, eds. Rosenberg, R. N. & Harding, A. E. (Butterworths, London), pp. 22–34.
4. Kaplan, B. & Finch, C. (1982) in *Molecular Approaches to Neurobiology*, ed. Brown, I. (Academic, New York), pp. 71–98.
5. Snider, B. J. & Morrison-Bogorad, M. (1992) *Brain Res. Rev.* **17**, 263–282.
6. Weissman, S. M. (1987) *Mol. Biol. Med.* **4**, 133–143.
7. Ko, M. S. H. (1990) *Nucleic Acids Res.* **18**, 5705–5711.
8. Patanjali, S. R., Parimoo, S. & Weissman, S. M. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 1943–1947.
9. Soares, M. B. (1994) in *Automated DNA Sequencing and Analysis Techniques*, ed. Venter, J. C. (Academic, London), pp. 110–114.
10. Adams, M. D., Soares, M. B., Kerlavage, A. R., Fields, C. & Venter, J. C. (1993) *Nat. Genet.* **4**, 373–380.
11. Dente, L., Cesareni, G. & Cortese, R. (1983) *Nucleic Acids Res.* **11**, 1645–1655.
12. Vieira, J. & Messing, J. (1987) *Methods Enzymol.* **153**, 3–11.
13. Rubenstein, J. L. R., Brice, A. E. J., Ciaranello, R. D., Denney, D., Porteus, M. H. & Usdin, T. B. (1990) *Nucleic Acids Res.* **18**, 4833–4842.
14. Grunstein, M. & Hogness, D. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 3961–3965.
15. Feinberg, A. P. & Vogelstein, B. (1983) *Anal. Biochem.* **132**, 6–13.
16. Feinberg, A. P. & Vogelstein, B. (1984) *Anal. Biochem.* **137**, 266–267.
17. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403–410.
18. Khan, A. S., Wilcox, A. S., Polymeropoulos, M. H., Hopkins, J. A., Stevens, T. J., Robinson, M., Orpana, A. K. & Sikela, J. M. (1992) *Nat. Genet.* **2**, 180–185.
19. Hawkins, J. D. (1988) *Nucleic Acids Res.* **16**, 9893–9908.
20. Hahn, W. E., Van Ness, J. & Maxwell, I. H. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 5544–5547.
21. Pearson, W. R., Davidson, E. H. & Britten, R. J. (1977) *Nucleic Acids Res.* **4**, 1727–1737.
22. Van Ness, J. & Hahn, W. E. (1982) *Nucleic Acids Res.* **10**, 8061–8077.
23. Chaudhari, N. & Hahn, W. E. (1983) *Science* **220**, 924–928.
24. Bantle, J. A. & Hahn, W. E. (1976) *Cell* **8**, 139–150.
25. Grouse, L. D., Schrier, B. K., Bennett, E. L., Rosenzweig, M. R. & Nelson, P. G. (1978) *J. Neurochem.* **30**, 191–203.
26. Milner, R. J. & Sutcliffe, J. G. (1983) *Nucleic Acids Res.* **11**, 5497–5520.
27. Galau, G. A., Britten, R. J. & Davidson, E. H. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 1020–1023.
28. Welsh, J., Liu, J.-P. & Efstratiadis, A. (1990) *Genet. Anal. Technol. Appl.* **7**, 5–17.
29. Mandel, P., Rein, H., Harth-Edel, S. & Mardell, R. (1964) in *Comparative Neurochemistry*, ed. Richter, D. (Macmillan, New York), pp. 149–163.
30. Winick, M. (1968) *Pediatr. Res.* **2**, 352–355.
31. Mazo, A. M., Minchenko, A. G., Avdonina, T. A., Gause, G. G. & Pusyriov, A. T. (1983) *Mol. Biol. Rep.* **9**, 155–161.
32. Baserga, S. J., Linnenbach, A. J., Malcolm, S., Ghosh, P., Malcolm, A. D. B., Takeshita, K., Forget, B. G. & Benz, E. J., Jr. (1985) *Gene* **35**, 305–312.
33. Christianson, T. W. & Clayton, D. A. (1988) *Mol. Cell. Biol.* **8**, 4502–4509.
34. Sasaki, Y. F., Ayusawa, D. & Oishi, M. (1994) *Nucleic Acids Res.* **22**, 987–992.



2

Molecular Cloning

A LABORATORY MANUAL

Sambrook Fritsch Maniatis

Abundant mRNAs

Initially, cDNA cloning was used to obtain copies of abundant mRNAs such as those encoding globin, immunoglobulins, and ovalbumin. In these cases, the RNA species of interest constitutes as much as 50–90% of the total poly(A)$^+$ cytoplasmic RNA isolated from specific types of differentiated cells. Consequently, no further purification of the particular mRNA is required before double-stranded cDNA is synthesized and cloned. The desired cDNA clones can easily be identified by nucleic acid hybridization. The probes consist either of ^{32}P-labeled single-stranded cDNA synthesized in vitro by reverse transcriptase, using as the template mRNA preparations that are rich in the sequences of interest, or of mRNA that has been partially fragmented by limited alkaline hydrolysis and end-labeled by phosphorylation. As a good approximation, the mRNA sequences of interest will be represented in both the probe and the cloned double-stranded cDNAs in proportion to their abundances in the original preparation of mRNA. In cases such as globin, immunoglobulins, and ovalbumin, the chances are high that any colony hybridizing strongly to the probe will contain the desired DNA sequences. Although used extensively in the early days of cDNA cloning, this method no longer finds wide application, since few systems remain in which interesting uncloned mRNAs represent a sufficiently high proportion of the starting population.

Low-abundance mRNAs

mRNAs that represent less than 0.5% of the total mRNA population of the cell are classified as "low-abundance" or "rare" mRNAs. The isolation of cDNA clones for mRNAs of this type presents two major problems: (1) construction of a cDNA library whose size is sufficient to ensure that the clone of interest has a good chance of being represented and (2) identification and isolation of the clone(s) of interest.

Methods of Enrichment

A typical mammalian cell contains between 10,000 and 30,000 different mRNA sequences (Davidson 1976). Not all of these sequences are represented equally in the steady-state population of mRNA molecules. Instead, the proportional representation of each sequence depends on its rate of synthesis and half-life: Genes that are actively transcribed into stable mRNAs will make a greater contribution to the pool of mRNA molecules than genes that are transcribed sluggishly into less stable mRNAs. Williams (1981) has determined the number of clones necessary to construct a complete cDNA library from a human fibroblast cell that contains approximately 12,000 different mRNA sequences. Low-abundance mRNAs (<14 copies/cell) constitute approximately 30% of the mRNA, and there are about 11,000 different mRNAs belonging to this class. The minimum number of cDNA clones required to obtain a complete representation of mRNAs of this class is therefore $11{,}000/0.30 \cong 37{,}000$. Of course, because of sampling variation and/or preferential cloning of certain sequences, a much larger number of recombinants must be obtained to increase the chances that any given clone will be represented in the library. The number of clones required to achieve a given probability that a low-abundance mRNA will be present in a cDNA library is

$$N = \frac{\ln(1 - P)}{\ln(1 - 1/n)}$$

where N = the number of clones required, P = the probability desired (usually 0.99), and $1/n$ = the fractional proportion of the total mRNA that is represented by a single type of rare mRNA.

Therefore, to achieve a 99% probability of obtaining a cDNA clone of an mRNA present in human fibroblasts at a frequency of approximately 14 molecules/cell:

$P = 0.99$
$1/n = 1/37{,}000$
$N = 170{,}000$

Unfortunately, many mRNAs of interest are present at even lower levels (1 molecule/cell is not unusual [Toole et al. 1984; Wood et al. 1984]). Furthermore, it is often necessary to clone cDNAs from populations of mRNAs isolated from tissues that consist of several cell types. In such cases, the frequency at which the sequences of interest are represented in the initial preparation of mRNA may be reduced still further, and it then becomes necessary to construct and screen libraries that contain several million independent cDNA clones. During the last few years, the efficiency with which cDNA can be synthesized and cloned has increased to the point where cDNA libraries of this size can be generated routinely from 10 μg or less of poly(A)$^+$ mRNA. In principle, there is no a priori reason why even the most difficult cDNA clones—those corresponding to a very rare mRNA of large size—cannot be identified in such comprehensive libraries. However, screening large numbers of cDNA clones is both tedious and expensive. Methods have therefore been devised to enrich either the starting population of mRNA molecules or double-stranded cDNA synthesized from it for sequences of interest. Enrichment allows the size of the cDNA library to be reduced and decreases the cost and labor involved in screening for the desired cDNA clones.

It is difficult to offer specific guidelines regarding the circumstances that require enrichment procedures. As a rule of thumb, fractionation of mRNA is probably unnecessary if the cDNA of interest is expected to be present at a frequency $\geqslant 1$ in 10^6 in a library of cDNA clones synthesized from unfractionated mRNA. Enrichment becomes more attractive as the number of clones to be screened increases above one million. When designing a scheme to clone a specific cDNA, it is therefore important to know the approximate frequency with which the mRNA of interest occurs in the bulk, unfractionated population of mRNA molecules. In the absence of nucleic acid probes, an indirect method must be used to measure this frequency. Usually, the mRNA preparation is translated in a cell-free system and the total amount of radioactivity incorporated into protein is measured. The polypeptide of interest is then immunoprecipitated and identified by electrophoresis through an SDS-polyacrylamide gel. The amount of radioactivity in the excised band is then measured and used to calculate the proportion of the total counts that have been incorporated into the protein of interest. This proportion is taken as a measure of the frequency with which the mRNA occurs in the bulk population. Despite its obvious limitations, this method usually yields estimates that are sufficiently reliable to allow rational schemes for cDNA cloning to be devised.

Clearly, fractionation works best for mRNAs that are much larger or smaller in size than the bulk mRNA of the cell. The modal size of the mRNA population extracted from most types of mammalian cells is approximately 1.8 kb, and mRNAs smaller in size than 700 bases or larger than 4 kb can be enriched at least tenfold by a single round of density gradient centrifugation carried out under denaturing conditions. However, it is important to remember that it is not possible to predict with certainty the size of an mRNA from the size of a protein for which it codes. There is considerable variation in the sizes of the untranslated regions of mRNAs (particularly the 3′ untranslated regions); many proteins purified from cells are cleavage products of larger precursors and many undergo extensive posttranslational modification. However, the size of the unmodified polypeptide chain provides a minimal estimate of the size of the mRNA: 10,000 daltons of an average polypeptide is encoded by approximately 280 bases of mRNA.

FRACTIONATION OF mRNA BY SIZE

The simplest method to enrich preparations of mRNA for sequences of interest is to fractionate them according to size. Electrophoresis through agarose gels gives the best separation of molecules of mRNA of different sizes, but the recovery of RNA from gel slices is generally poor. Sedimentation through sucrose gradients formed in nondenaturing solvents results in good recovery, but the presence of secondary structure in the RNA often confounds effective fractionation. The method of choice, therefore, is sucrose gradient centrifugation in the presence of an agent, such as methylmercuric hydroxide, that denatures secondary structure in RNA (Schweinfest et al. 1982) (for experimental protocol, see Chapter 7, page 7.35). Each fraction is then assayed for the presence of the mRNA that codes for the relevant polypeptide. Typically, an aliquot of the RNA in each fraction is translated in a cell-free system and the resulting polypeptides are analyzed by immunoprecipitation and electrophoresis through polyacrylamide gels. Alternatively, aliquots are injected into *Xenopus* oocytes (for review, see Melton 1987) and the resulting products are assayed either for biological activity or by immunoprecipitation and gel electrophoresis. The fraction that directs the synthesis of the greatest amount of the polypeptide product is then used as the starting material for construction of a cDNA library.

FRACTIONATION OF cDNA

Until a few years ago, fractionation of mRNA was the method of choice for cloning of mRNAs that code for large proteins (e.g., rat skeletal muscle tropomyosin [Medford et al. 1980] and chick creatine kinase [Schweinfest et al. 1982]). However, as methods for the synthesis of cDNA have improved, fractionation of double-stranded cDNA has become a more practical alternative, and there are now many examples of extremely large cDNAs that have been cloned by fractionating cDNA rather than the mRNA from which it was copied (e.g., human factor VIII:C [Toole et al. 1984; Wood et al. 1984] and human sucrase-isomaltase [Hunziker et al. 1986]). Fractionation of cDNA has major advantages: DNA is less susceptible than mRNA to degradation by contaminating nucleases; it can be fractionated more accurately by electrophoresis through agarose gels; and, finally, since the fractionation can be carried out at a late stage during the cDNA cloning protocol, the chances of subsequent mishaps are reduced and the probability of obtaining a full-length clone of cDNA is increased. Fractionation is usually carried out after all of the enzymatic reactions involved in cDNA synthesis have been completed and just before the cDNA is inserted into a vector. In the detailed protocol described later in this chapter, fractionation is carried out after synthetic linkers, added to the termini of double-stranded cDNA, have been digested with a restriction enzyme. The cDNA is fractionated by electrophoresis through an agarose gel of appropriate porosity, using markers whose sizes are known accurately. Molecules of the desired size are recovered and inserted into the vector.

IMMUNOLOGICAL PURIFICATION OF POLYSOMES

An alternative method of enrichment involves the use of antibodies to purify polysomes that are synthesizing the polypeptide of interest. The technique described originally (Palacios et al. 1972; Schechter 1973), which involved the immunoprecipitation of polysomes, worked well for mRNAs encoding abundantly synthesized proteins such as albumin and immunoglobulin, although attempts to apply the method to mRNAs of lesser abundance were generally disappointing. However, the more recent use of immunoaffinity columns (Schutz et al. 1977) and protein A–Sepharose columns (Shapiro and Young 1981) has led to a resurgence of the technique. For example, Korman et al. (1982) used a monoclonal antibody directed against the heavy chain of the human HLA-DR histocompatibility antigen to bind polysomes synthesizing the nascent protein to protein A–Sepharose columns. The polysomes were then dissociated with EDTA and the mRNA isolated by oligo(dT) chromatography. The immunoaffinity-purified mRNA, which represented only 0.01–0.05% of the total mRNA, was used both to prepare cDNA probes and to construct cDNA clones. Using similar methods with polyclonal antisera, Kraus and Rosenberg (1982) obtained a 6300-fold purification of the mRNA that codes for rat liver cystathionine β-synthase and Russell et al. (1983) isolated cDNA clones for the bovine low-density-lipoprotein receptor, whose mRNA is present at approximately 80 copies per cell in bovine adrenal cells.

Although a powerful technique, immunoaffinity purification of polysomes cannot be applied universally. First, it clearly will not work unless a reliable source of material is available from which to isolate functional polysomes. This is not always possible, especially when the starting material is a tissue or organ that is not commonly available. Second, it has not yet been shown to work for mRNAs that are extremely rare (1 molecule/cell or less). Furthermore, the success of the method depends entirely on the specificity, avidity, and type of the particular antibody, and it is not always possible to translate the results obtained with one antibody directly to another. Finally, the method requires the use of relatively large quantities of antibody. Partly because of these difficulties, immunoaffinity purification of polysomes has been superseded by development of cDNA cloning vectors (e.g., λgt11 and λgt18–23) that allow the direct isolation of cDNA clones that code for specific antigens. Whether or not the method is used extensively in the future will depend on improving its sensitivity to the point where it provides significant enrichment of polysomes carrying extremely rare mRNAs.